Abstract Title Page
Not included in page count.

Title: Combining propensity score methods and complex survey data to estimate population treatment effects

Authors and Affiliations:

Elizabeth A. Stuart
Departments of Mental Health, Biostatistics, and Health Policy and Management
Johns Hopkins Bloomberg School of Public Health
estuart@jhu.edu

Nianbo Dong
Department of Educational, School, and Counseling Psychology
College of Education
University of Missouri
dongn@missouri.edu

David Lenis
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
dlenis@jhsph.edu

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Large scale, complex survey designs are widely used in educational research. These surveys, such as the Early Childhood Longitudinal Study Birth Cohort (ECLS-B) and the Schools and Staffing Survey (SASS), aim to collect information on a sample that is formally representative of its target population, and they do this using a sampling frame and sample design. In large-scale surveys, the sample design may be complicated and the resulting sampling weights for each individual reflect these varying probabilities of selection. Those sampling weights then need to be accounted for in analyses in order to draw inferences relevant to the target population. (see, e.g., Hansen, Madow and Tepping, 1983; Korn & Graubard, 1995a, 1995b).

Complex surveys are often used to estimate causal effects regarding the effects of interventions or exposures of interest. Propensity scores (Rosenbaum & Rubin, 1983) have emerged as one popular and effective tool for causal inference in non-experimental studies, as they can help ensure that groups being compared are similar with respect to a large set of observed characteristics. However, little work has investigated how best to combine propensity scores and complex survey data to estimate population treatment effects. This has led to confusion in the literature, with many applied researchers using inappropriate methods or claiming representativeness of study results when the analysis does not warrant such claims (DuGoff, Schuler and Stuart, 2014; Ridgeway, Kovalchik, Griffin, and Kabeto, 2015). One way to think about the complication of estimating population treatment effects using data from a complex survey is that when there are heterogeneities in treatment assignment, sample selection probabilities, and treatment effects, failure to take into account sampling weights might cause biased population treatment effect estimates. Ignoring sampling weights leads mainly to external validity bias, which occurs when people inappropriately make inferences from the unrepresentative analytic sample to the target population. This work aims to clarify the results and recommendations regarding the use of propensity scores with complex survey data.

Purpose / Objective / Research Question / Focus of Study:

This presentation will address the need for clear specification of the estimand of interest when estimating causal effects with data from complex surveys, and methods questions such as whether the survey weights need to be included in the propensity score model, and how to combine the propensity score approach and methods for complex survey data when comparing outcomes between treated and control groups. In particular, the work will distinguish estimands of the "Sample Average Treatment Effect" from the "Population Average Treatment Effect" (and the related quantities, the "Sample Average Treatment Effect on the Treated" and "Population Average Treatment Effect on the Treated") and clarify which propensity score methods estimate which quantity, as well as which propensity score methods work best when used with complex survey data.

Population / Participants / Subjects:

We will use simulated data to investigate the performance of statistical methods, and will also illustrate the use of the methods using data from the ECLS-B. The ECLS-B used a complex survey design that randomly sampled 14,000 children born in 2001 in the United States, producing a nationally representative sample. Of the 14,000 children selected for study participation, approximately 10,700 of the children who participated in the first round of data collection (at 9 months of age) constitute the baseline sample for this study.

Significance / Novelty of study:

Many education research studies aim to estimate causal effects using data from complex surveys. Propensity scores are an effective approach for estimating causal effects in non-experimental studies. However, to this point there has not been good guidance regarding how to use propensity score methods with complex survey data, leading to a landscape of researchers sometimes using approaches that are not appropriate or claiming representativeness of study findings when that is not warranted. This work will help clarify the approaches that will yield accurate estimates of treatment effects when using propensity score methods with complex survey data. The purpose of this study is to systematically investigate the appropriate use of sampling weights in propensity score analysis.

Statistical, Measurement, or Econometric Model:

This work investigates a number of statistical approaches for using propensity scores with complex survey data, focusing both on propensity score estimation (Stage1) and use (Stage 2). For Stage 1, there are three options for handling sampling weights in estimating propensity scores: (1) "NoWt": ignoring sampling weights, (2) "WtCov": using the sampling weights as a covariate, and (3) "WtModel": estimating a weighted logistic regression model using the sampling weights. We consider five propensity score methods for Stage 2: (1) "Cov": using the propensity score as a covariate, (2) "Match": matching (1-to-1, 1-to-N, etc.), (3) "Strat"; stratification (also called subclassification) which includes conventional stratification (Rosenbaum and Rubin, 1985) as well as (4) "MMWS": marginal mean weighting through propensity score stratification (see Hong, 2010 and Posner and Ash, 2009), and (5) "Wt": weighting, using the inverse of the probability of treatment. In these Stage 2 approaches the options for incorporating survey weights depend on which propensity score method is to be used and on which effect (PATT or PATE) is of interest, but broadly the options are: (1) "NoWt": ignoring sampling weights, (2) "WtModel": estimating a weighted regression model using the original sampling weights, and (3) "RWtModel": estimating a weighted regression model using the reweighted sampling weights. Table 1 summarizes the options of handling sampling weights in propensity score analyses that were identified in literature or with natural extension.

Usefulness / Applicability of Method:

The methods are illustrated using data from the ECLS-B, estimating the effect of a child care subsidy on child math achievement at kindergarten. Estimates of the Sample Average Treatment Effect and the Population Average Treatment Effect are presented to help audience members

understand the distinction between those quantities. Furthermore, informed by the simulation study, we provide guidelines on how to combine the propensity score approach and methods for complex survey data when comparing outcomes between treated and control groups.

Research Design:

Guided by DuGoff, Schuler and Stuart (2014), Ridgeway, Kovalchik, Griffin, and Kabeto, (2015), and our recent theoretical work on this topic, we use Monte Carlo simulations to evaluate the performance of the methods described above in terms of bias and MSE under various situations when there are heterogeneities in treatment assignment and in treatment effects in the population. We assume that the population consists of ten strata that are heterogeneous in treatment assignment, treatment effects, and sampling weights. For treatment assignment, we adopted Hong's (2010) simulation framework with significant extensions to fit our study design. Hong (2010), and we, consider three true binary treatment assignment models and two true outcome models, with a single normally distributed covariate, X. In addition, we allow the parameters for treatment assignment and treatment effects to vary across ten strata. We assume that the total size of the population is 20,000, divided into 10 strata of size 2,000 each, and that a stratified sampling framework is used to select a sample with different selection possibilities (inverse of the sampling weights) from each stratum. The strata are defined independent of X. The sampling probabilities vary from 1 to 1/20 (.05), with the resulting weights varying from 1 to 20, across the 10 strata.

For each combination of treatment assignment and outcome model we generate a population with our designed parameters. We then draw a random sample from this population using the stratified sampling design described above. We repeat the above process 1000 times to generate 1000 simulated datasets. For all propensity score methods, we use the mis-specified propensity score model to estimate the propensity score, weighted by the sampling weights for some of the methods. We then use various propensity score methods to estimate the treatment effects. For simplicity and to enable clear comparisons among the propensity score methods, we do not include the covariate in the outcome analysis to remove the effects of covariate in reducing treatment effect estimate bias. We compare the five propensity score methods described above, and the various combinations of using (or not using) the survey weights. For covariate adjustment, matching, stratification, and MMWS we use the logit of the propensity score. For propensity score weighting, the weights are calculated using the propensity score itself. For each of the 1000 random samples we obtain population effect estimates using each of the five methods and their variations, we calculate the bias and mean square error (MSE) across samples to evaluate the performance of these methods. The true PATE and PATT are known, as calculated using the potential outcomes in the population.

Findings / Results:

Table 2 provides an example of the results obtained under one particular simulation setting. We find that when survey weights are ignored in analyses misleading conclusions regarding population treatment effects may be drawn. However, accurate results can be obtained if the sampling weights are taken into account in the outcome analyses. We found that the relatively

simple approach of multiplying propensity score weights by the survey weights can work well, but is sensitive to misspecification of the propensity score model.

For the ECLS-B application, Table 3 presents the weighted sample means of covariates for: (1) the full sample of subsidy recipients, (2) the full sample of subsidy non-recipients, (3) the matched subsidy non-recipients (resulting from 1-to-1 optimal matching on the logit of the propensity score estimated from an unweighted logistic regression model, "Match-NoWt-RWtModel"), and (4) the matched subsidy non-recipients (resulting from 1-to-1 optimal matching on the logit of the propensity score estimated from a weighted logistic regression model, "Match-WtModel-RWtModel"). Using Match-WtModel-RWtModel, only one covariate with standardized bias bigger than 0.25, Column 7 in Table 3), which resulted in slightly better performance in balancing covariates than when the propensity score estimation did not use the survey weights ("Match-NoWt-RWtModel"); for that approach three covariates had standardized biases bigger than 0.25, (Column 6 in Table 3). However, both approaches improved balance relative to the unmatched sample, in which 11 of the 18 covariates had standardized biases larger than 0.25 (Column 5 in Table 3).

Table 4 presents a summary of the effects of the subsidy on children's kindergarten math score. The PATE and PATT estimates varied quite a bit across the propensity score methods. The majority of analyses suggested a significant negative effect of subsidy receipt on child kindergarten math scores for the population represented by the ECLS-B.

Conclusions:

There are several limitations in this paper and many directions for further research in this area. For example, both this paper and DuGoff, Schuler and Stuart (2014) used only one covariate in the simulations. Future studies could conduct simulations using multiple covariates and a data structure from a real survey. The current paper also was primarily concerned with sampling weights and did not take account other survey features, such as clustering or strata. The simulations generated the weights using a stratified sampling design; more complicated survey designs should be considered in future research. In addition, both DuGoff, Schuler, and Stuart (2014) and the current study concerned normally distributed outcome variables. Other outcome distributions should be examined in future studies. The current study also was only concerned with the point estimate of the population treatment effect and estimated the standard error based on the outcome models approach (e.g., weighted regression models; model-based inference), rather than also investigating variance estimates. Furthermore, it was a challenge to model treatment effect heterogeneity in the simulation; future studies could calibrate the extent of treatment effect heterogeneity better to reflect real-word scenarios.

In conclusion, it is important for researchers to think carefully about their estimand of interest, and use methods appropriate for that estimand. If interest is in drawing inferences to the survey target population (i.e., in estimating the PATE or PATT) it is important to take the survey weights into account, particularly in the outcome analysis stage. We hope that this paper raises awareness of this issue and provides a caution, as well as concrete suggestions, for researchers interested in examining causal effects in populations represented by sample surveys.

Appendices

Not included in page count.

Appendix A. References

DuGoff, E. H, Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: applying propensity score methods to complex surveys. *Health Services Research, 49*(1):284-303. doi: 10.1111/1475-6773.12090.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, *Vol. 78*, No. 384, pp. 776-793

Hong, G. (2010). Bias in multilevel data marginal mean weighting through stratification: Adjustment for selection. *Journal of Educational and Behavioral Statistics, 35* (5), 499–531. doi: 10.3102/1076998609359785

Korn, E. L. & Graubard, B. I. (1995a). Analysis of large health surveys: Accounting for the sampling designs. *Journal of the Royal Statistical Society*, *158*, pp. 263–295.

Korn, E. L. & Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49* (3), pp. 291–295. doi:10.1080/00031305.1995.10476167

Posner, M. A., & Ash, A. S. (2009). Comparing Weighting Methods in Propensity Score Analysis, Retrieved March 11, 2011, from http://www.stat.columbia.edu/~gelman/stuff_for_blog/posner.pdf

Reardon, S. F., Cheadle, J. E. & Robinson, J. P. (2009). The Effect of Catholic Schooling on Math and Reading Development in Kindergarten Through Fifth Grade. *Journal of Research on Educational Effectiveness, 2*:1, 45-87

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity Score Analysis with Survey Weighted Data. Journal of Causal Inference, 3(2), 237 – 249. DOI 10.1515/jci-2014-0039

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*:41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist. 39* 33–38.

Schonlau, M., Soest, A. V., Kapteyn, A., Couper, M., & Winter, J. (2004). Adjusting for selection bias in Web surveys using propensity scores: the case of the Health and Retirement Study. In Proceedings of the Section on Survey Research Methods. Retrieved

on September 6, 2012 on
http://www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000032.pdf

Appendix B. Tables and Figures

Tables and Figures

Table 1: Summary of Options of Handling Sampling Weights in Propensity Score Analyses

Propensity Score Methods	Ways of Handling Sampling Weights at Two Stages of Propensity Score Methods			
	Label	Stage 1: Estimation (Estimating Propensity Scores)	Stage 2: Use (Outcome Analysis)	
			PATE	PATT
1. Using propensity scores as covariate	Cov-NoWt-NoWt	Ignoring	Ignoring	NA
	Cov-NoWt-WtModel	Ignoring	Weighted	NA
	Cov-WtModel-WtModel	Weighted	Weighted	NA
	Cov-WtCov-NoWt	Covariate	Ignoring	NA
2. Matching	Match-NoWt-NoWt	Ignoring	NA	Ignoring
	Match-NoWt-RWtModel	Ignoring	NA	Reweight 1
	Match-WtModel-RWtModel	Weighted	NA	Reweight 1
	Match-WtCov-NoWt	Covariate	NA	Ignoring
3a. Conventional Stratification/ Subclassification	Strat-NoWt-NoWt	Ignoring	Ignoring	Ignoring
	Strat-NoWt-WtModel	Ignoring	Weighted	Weighted
	Strat-WtModel-WtModel	Weighted	Weighted	Weighted
	Strat-WtCov-NoWt	Covariate	Ignoring	Ignoring
3b. Marginal Mean Weighting through Propensity Score Stratification	MMWS-NoWt-NoWt	Ignoring	Ignoring	Ignoring
	MMWS-NoWt-WtModel	Ignoring	Weighted	Weighted
	MMWS-WtModel-WtModel	Weighted	Weighted	Weighted
	MMWS-WtCov-NoWt	Covariate	Ignoring	Ignoring
4. Weighting	Wt-NoWt-NoWt	Ignoring	Ignoring[a]	Ignoring[a]
	Wt-NoWt-RWtModel	Ignoring	Reweight 2	Reweight 2
	Wt-WtModel-RWtModel	Weighted	Reweight 2	Reweight 2
	Wt-WtCov-NoWt	Covariate	Ignoring[a]	Ignoring[a]

[a]Ignoring survey sampling weights but still using the propensity score weights in weighted regression models to estimate PATE and PATT.
"Reweight 1" refers to using the original sampling weights for the matched (treatment) sample and reweighting the matching (control) sample to keep the summed weights equal for matched sample and matching sample and the new weights for the matching sample proportional to their original sampling weights, then using weighted analysis to estimate ATT (Reardon, Cheadle, & Robinson, 2009).
"Reweight 2" refers to using the product of the original sampling weights and propensity score weights as new weights in weighted analysis (Schonlau et al, 2004).

Table 2: Bias and Mean Square Error (MSE) Using the Sample in Simulation 1

Propensity Score Methods	Ways of Handling Sampling Weights at Two Stages of Propensity Score Methods	Normal, linear outcome				Normal, nonlinear outcome			
		PATE		PATT		PATE		PATT	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
1. Using propensity scores as covariate	Cov-NoWt-NoWt	-0.350	0.123	NA	NA	-0.349	0.122	NA	NA
	Cov-NoWt-WtModel	0.014	0.002	NA	NA	0.014	0.002	NA	NA
	Cov-WtModel-WtModel	0.014	0.002	NA	NA	0.014	0.002	NA	NA
	Cov-WtCov-NoWt	-0.350	0.123	NA	NA	-0.349	0.122	NA	NA
2. Matching (1-to-1 greedy matching)	Match-NoWt-NoWt	NA	NA	-0.455	0.207	NA	NA	-0.423	0.179
	Match-NoWt-RWtModel	NA	NA	-0.103	0.013	NA	NA	-0.070	0.007
	Match-WtModel-RWtModel	NA	NA	-0.081	0.008	NA	NA	-0.053	0.005
	Match-WtCov-NoWt	NA	NA	-0.455	0.207	NA	NA	-0.423	0.180
3a. Conventional Stratification/ Subclassification	Strat-NoWt-NoWt	-0.369	0.137	-0.356	0.127	-0.365	0.133	-0.352	0.124
	Strat-NoWt-WtModel	0.001	0.002	-0.019	0.002	0.007	0.002	-0.014	0.002
	Strat-WtModel-WtModel	0.001	0.002	-0.019	0.002	0.007	0.002	-0.014	0.002
	Strat-WtCov-NoWt	-0.369	0.136	-0.356	0.127	-0.364	0.133	-0.351	0.124
3b. Marginal Mean Weighting through Propensity Score Stratification	MMWS-NoWt-NoWt	-0.369	0.137	-0.356	0.127	-0.365	0.133	-0.352	0.124
	MMWS-NoWt-WtModel	0.001	0.002	-0.019	0.002	0.007	0.002	-0.014	0.002
	MMWS-WtModel-WtModel	0.001	0.002	-0.019	0.002	0.007	0.002	-0.014	0.002
	MMWS-WtCov-NoWt	-0.369	0.136	-0.356	0.127	-0.364	0.133	-0.351	0.124
4. Weighting	Wt-NoWt-NoWt	-0.259	0.068	-0.255	0.066	-0.285	0.082	-0.229	0.054
	Wt-NoWt-RWtModel	0.046	0.006	0.031	0.004	0.038	0.005	0.067	0.008
	Wt-WtModel-RWtModel	0.146	0.024	0.113	0.016	0.110	0.015	0.145	0.026
	Wt-WtCov-NoWt	-0.259	0.067	-0.255	0.066	-0.285	0.082	-0.229	0.054

Note: The PATE and PATT for the population are both 2.30, and the PATE and PATT for the sample ignoring sampling weights are both 1.96. The theoretical values of biases that use the sample ignoring sampling weights to estimate the population PATE and PATT are both -0.34.

Table 3: Covariate Balance Checking Based on Weighted Sample Means

Variable	Sampling Weighted Means				Standardized Bias Based on Weighted Means		
	(1) Full sample of subsidy recipients	(2) Full sample of subsidy non-recipients	(3) Matched subsidy non-recipients (Match-NoWt-RWtModel)	(4) Matched subsidy non-recipients (Match-WtModel-RWtModel)	(5) Standardized Bias: (1)-(2)	(6) Standardized Bias: (1)-(3)	(7) Standardized Bias: (1)-(4)
Pre-test measures of cognitive skills							
Preschool math score	26.71	29.65	26.93	26.71	-0.30	-0.03	0.00
Preschool reading score	22.10	25.54	23.29	22.64	-0.33	-0.13	-0.06
2 years mental score	126.94	127.99	123.77	126.17	-0.10	0.33	0.08
9 months mental score	77.14	77.02	73.67	76.23	0.01	0.35	0.09
Child characteristics							
Birth weight[a]	3.17	3.33	2.86	2.88	-0.19	0.35	0.34
Percent female	50	50	56	49	0.00	-0.14	0.01
Percent black[b]	37	13	35	41	0.66	0.05	-0.08
Percent Hispanic[b]	21	22	25	19	-0.03	-0.10	0.04
Percent other race[b]	9	7	16	19	0.05	-0.19	-0.27
Percent started K in 2007	32	26	21	25	0.15	0.27	0.18
Age (months) of K assessment	68.27	68.15	67.84	67.50	0.03	0.10	0.18
Family characteristics at age 2							
Family income[c]	2.68	5.61	3.07	2.85	-0.63	-0.12	-0.05
Percent welfare recipients	0.27	0.06	0.27	0.19	0.79	-0.02	0.18
Percent WIC recipients	0.67	0.38	0.75	0.67	0.60	-0.19	-0.01
Percent single parents	0.58	0.19	0.52	0.56	1.02	0.12	0.04
Mother's years of education	12.54	13.46	12.65	12.86	-0.35	-0.06	-0.15
Mother's age at child's birth	23.94	27.53	24.96	24.19	-0.57	-0.16	-0.04
Percent speak English at home	9	16	9	12	-0.17	-0.02	-0.10
N	250	5400	250	250			

Source: Early Childhood Longitudinal Study – Birth cohort

Note: In compliance with NCES policy, all sample sizes have been rounded to the nearest 50.

[a] Birthweight is measured in 1,000 gram units; [b] referent group is white; [c] income is measured in $10,000 units

Table 4: Summary of the Effects of Subsidy on Kindergarten Math

Regression/ Propensity Score Methods	Ways of Handling Sampling Weights at Two Stages of Propensity Score Methods				Subsidy Effect[a]		Sample Size
	Label	Stage 1: Estimation (Estimating Propensity Scores)	Stage 2: Use (Outcome Analysis)				
			PATE	PATT	PATE	PATT	
Replication of Hawkinson, Griffen, Dong, & Maynard (2013)							
OLS Regression		NA	Weighted	NA	-1.61 (.45)	NA	5,650
Matching (1-to-1 optimal matching based on propensity score probability)		Ignoring	NA	Ignoring	NA	-1.74 (.59)	500
OLS Regression		NA	Ignoring	NA	-1.23 (.44)	NA	5,650
Matching (1-to-1 optimal matching based on propensity score logit)	Match-NoWt-NoWt	Ignoring	NA	Ignoring	NA	-0.16^ (.59)	500
	Match-NoWt-RWtModel	Ignoring	NA	Reweight 1	NA	0.42^ (.61)	500
	Match-WtModel-RWtModel	Weighted	NA	Reweight 1	NA	-1.58 (.58)	500
Stratification/ Subclassification on propensity score logit	Strat-NoWt-NoWt	Ignoring	Ignoring	Ignoring	-1.37 (.48)	-1.28 (.47)	5,650
	Strat-NoWt-WtModel	Ignoring	Weighted	Weighted	-1.41 (.49)	-1.27 (.46)	5,650
	Strat-WtModel-WtModel	Weighted	Weighted	Weighted	-1.66 (.51)	-1.30 (.48)	5,650
Weighting	Wt-NoWt-NoWt	Ignoring	Ignoring 2	Ignoring 2	-0.82 (.17)	-1.27 (.18)	5,650
	Wt-NoWt-RWtModel	Ignoring	Reweight 2	Reweight 2	-0.62 (.17)	-1.62 (.17)	5,650
	Wt-WtModel-RWtModel	Weighted	Reweight 2	Reweight 2	-0.66 (.17)	-1.78 (.17)	5,650

Source: Early Childhood Longitudinal Study – Birth cohort
Note: In compliance with NCES policy, all sample sizes have been rounded to the nearest 50.
[a]All the other subsidy effect are statistically significant at alpha = 0.05 except the noted ones.
^$p > 0.05$